DOCUMENT RESUME

ED 151 950 EA 010 447

AUTHOR Nash, Nicholas
TITLE Delphi and Educational Research: A Review.
PUB DATE [78]
NOTE 55p.

EDRS PRICE MF-$0.83 HC-$3.50 Plus Postage.
DESCRIPTORS Educational Research; Futures (of Society); Higher Education; *Prediction; *Research Methodology; Research Tools
IDENTIFIERS *Delphi Technique

ABSTRACT

The purpose of this paper is to review the development of the Delphi technique and to explore some examples of its application in educational research. The discussion of the technique's development is intended to help potential users of the method see more clearly its liabilities as well as its potential. The eight examples of the technique's use cover three areas--educational goals and objectives, curriculum and campus planning, and effectiveness and cost-benefit measures. Criticism of Delphi as a research methodology is included. Despite these criticisms, it is suggested that Delphi may be useful for two reasons--it is a convenient and relatively nontechnical methodology that is appropriate for use with a population not primarily research oriented, and it appears a convenient vehicle for exploring probable and possible futures in what have been termed "squishy" areas. A bibliography is included. (Author/IRT)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED151950

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Delphi and Educational Research: A Review

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Nicholas Nash

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

Nicholas Nash
University of Minnesota[1]

EA 010 447

[1] Presently Associate Director, The University Council for Educational Administration, Columbus, Ohio.

In its purest form the Delphi Methodology is no more than a group of "experts" generating a number of opinions on one or more topics, and these opinions are refined over several rounds. The process of refining typically involves each expert's receiving the range of group opinions, expressed statistically or literally, and employing that information in re-estimating his/her own responses to the opinion items.

Recently, Delphi has become a relatively common vehicle for educational research, particularly at the policy level. The purpose of this paper is first to review Delphi's history and then to explore some examples of its application in educational research. By developing a sense of the context in which Delphi has evolved, potential users of the methodology may see more clearly its liabilities as well as its potential.

In Beckett's "Waiting for Godot" (1970), the following exchange occurs:

Vladimir: Do you recognize us?

Pozzo: I am blind.

Silence

Estragon: Perhaps he can see into the future.

Estragon is like many people whose model of the future presupposes that only mystics, necromancers, and the blind can prophesy through some

inexplicable power.[1] Implicit in this belief is the view that men and women who do not have this capacity have no window on tomorrow, that some hidden power is required to prophesy, and that tomorrow is, in some sense, not amenable to the direct influence of today's events.

At the very least, prediction is an art. Beyond whatever trappings in which the context of a prediction is cloaked, statements of future events are individually created symbols of an alluring or repugnant unknown, and they rely on some intuitive hunch, the assumptions for which may be based in knowledge, experience, fantasy, wishes, distortions, and so on. The measure of the accuracy of any prediction is whether the event occurs. On the other hand, the popular literature is replete with prognosticators of various kinds; their predictive accuracy may be almost secondary to the popular fascination with whatever the prognostications may be.

This interest in the artistic framework surrounding prediction contrasts with the human tendency to believe and act though tomorrow will be very much like today. The supposed symmetry of today and tomorrow minimizes uncertainty, for, as de Jouvenel observes, "the vast majority of our actions--at present, just as in the distant past--conform to recipes" (1967:9).

Scientific predictions may be almost a contradiction in terms. Polak commented:

---

[1]The Blind seer or soothsayer is a well-known literary figure--as in Shakespeare's Julius Caeser or Oedipus Rex by Sophocles.

> Both types of modern socio-scientific theory [classical natural science and empirical micro-sociological science] suffer from one and the same fault. They have removed the time dimension. . . . Both have unanimously excised the future as an unclean and infected thing. (1971:149)

One cannot know the future scientifically except (a) as one accepts the timelessness of relatively microcosmic principles or laws which are known, and (b) as one assumes that those laws will remain until better ones are discovered, tested, and assimilated. One cannot "know" the future except by presuming continuity.

Nisbet would shun any approximation to science, or even the guise of science, in predicting the future.

> What is in question, however, is whether the marvels of electronic technology, cybernetics, linear programing, systems analysis, game theory, and the like do add anything, can add anything to our success in an enterprise that is at least two hundred years old in the best: the serious, conscious business of predicting the future by observation of real or imagined continuities of event, change, and circumstance in time. (1968:61)

For Nisbet, modern methods cannot catch up in their net "the Random Event, the Maniac, the Prophet, and the Genius" (1968:66), and it is these factors which destroy the probabilities of ever predicting the future. He concludes that attempts to deal with the consequences of other kinds of

change really are not worthwhile.

The pseudo-science of forecasting is no better than the art of forecasting. de Jouvenel's classic exploration in The Art of Conjecture of these and other questions led him to a view not unsympathetic to Nisbet's in that he built no insurmountable walls between art and science, but was clearly more comfortable with art. The forecaster has two tasks, according to de Jouvenel:

> to consider what changes are likely to be produced by dominating processes, seen as strong and powerful tendencies during the period under consideration. (1967:114) The second . . . duty . . . is to combat the general feeling of uncertainty which the rapidity of change sheds indistinctly over all institutions. (1967:247)

If to forecast a decision concerning present values, scientific approaches appear contradictory or plainly unappealing, and pure artistry seems appealing but perhaps a bit unwise, then, it is possible (although not necessarily desirable) to search for ways to make art more scientific. By constructing a relatively consistent and durable frame around the artistic processes, perhaps forecasting might be more congenial to measurement, evaluation, replication, and prediction concerning the process of forecasting.

One noted attempt to construct the underpinnings for such a frame came in Helmer and Rescher's "On the Epistemology of the Inexact Sciences" in which it was argued that the differences between the inexact and exact sciences are a "fiction" and only "of degree" (1959:25). Helmer and Rescher believe that "a discipline which provides predictions of a less

precise character, but makes them correctly and in a systematic and reasoned way, must be classified as a science" (1959:25) [emphasis added].

If correctness accompanies systematic impression, then one may needs suffer a host of afflictions under the guise of science. Constraints are in order.

> An explanation . . . must beyond reasonable doubt establish its hypothesis as more credible than its negation. Of a prediction, on the other hand, we need to require only that it establish its hypothesis simply as more credible than any comparable alternative. (1959:32)

> Generally speaking, in any field in which our ability to forecast with precision is very limited, our actions of necessity are guided by only slight differences in the probability which we attach to possible future alternative states of the world, and consequently we must permit predictions to be based on far weaker evidence than explanations. (1959:32)

A third constraint focuses on the use of experts who are "rational . . . have a large background knowledge in that field, and whose predictions . . . with regard to hypotheses . . . in that field show a record of comparative success in the long run" (1959:36).

The imposition of antecedent criteria is intended to compensate in these inexact sciences for the methodological rigor of the exact sciences. Whether it does has been a matter of considerable interest. It is clear that considerable effort has been spent on trying to develop more effective

means of forecasting, means that, although they do not enjoy the profile of science, suggest an acceptable silhouette of it.

The apparent inappropriateness of the classical scientific model to forecasting future events mandates that another kind or other kinds of models/methods emerge, and in fact, many have. The range of techniques is prodigious. Ayres (1969), Johnston (1972), Martino (1972a, 1972b), and Jantsch (1967) offered extensive discussions of these various methods, including Delphi.

Dalkey suggested bluntly the necessity for a different model:

> We do not have neatly packaged theories, or formulae in hard books, for generating synthetic histories ahead of time. (1968b:2)

In much the same vein Helmer observed that "in dealing with social situations, there is not even a vestige of a mathematical model that might serve as a crutch for generalizations, and forecasting becomes entirely a matter of intuitive estimate." (1970:2) (See also Helmer, 1967a.)

The Delphi model, then, comes to rely on intuition in making future estimates and must "make the most of it by using only the judgment of experts specialized in the areas relevant to the subject matter of the desired forecast and by combining the opinions of the various experts as effectively as possible" (Helmer, 1970:2).

The pertinent assumptions underlying the intuitive approach of experts assume that "Some past pattern will repeat, or a present pattern will continue" (Martino, 1972b:20-21), and that by virtue of their rationality in combination with their knowledge, experts will be able to generate successful predictions.

These assumptions are of critical importance and are explored subsequently. What is of moment at this point is the link between knowledge and expertness. Helmer and Rescher wrote,

> This non-explicitness of background knowledge, which nonetheless may be significant or even predominantly important, is typical of the inexact sciences, as is the uncertainty as to the evidential weight to be accorded various pieces of prima facie information in view of indirect evidence provided by underlying regularities. Hence the great importance which must be attached to experts and to expertise in these fields. For the expert has at his ready disposal a large store of (mostly inarticulated) background knowledge and a refined sensitivity to its relevance, through the intuitive application of which he is often able to produce trustworthy personal probabilities regarding hypotheses in his area of expertness. (1959:38)

This summation of the rationale for Delphi is reasonable, and it is a rationale that is necessary since if it is not acceptable, then in Helmer and Rescher's words, "we must either for the present renounce all claims to systematized prediction in the social sciences, or . . . turn to unorthodox methods which are based upon judicious and systematic reliance on expert judgment" (1959:39).

The utilization of experts can be useful in many areas but, according to Martino, especially

> in fields that are poorly structured; in fields that are too new to have adquate [sic] historical data for the use of other methods; in fields where progress may be more dependent on external social and economic factors than on the technological factors intrinsic to the field; and in fields where ethical or moral considerations may weigh heavily. (1972b:22)

The broad areas where the systematically collected opinions of experts may be utilized are two: the softer policy and value sector and the harder technological one in which "the future characteristics of useful machines, procedures, or techniques" may be explored (Martino 1972b:2). This distinction is useful; although the buttressing assumptions for the method remain the same for each sector perused, the nature of the "technology" may--and does--have an impact on the process of estimation itself or the results of that estimation.

Weaver (1969) elucidated two trends in the investigation of thought and the future: the first trend concerned the exploration of the future time perspective and traits of the individual which, according to Weaver, came out of abnormal psychology; and second related to forecasting and public opinion literature with an emphasis on accuracy.

A third trend which seems to have been an influence on the evolution of Delphi is the individual and group problem-solving literature. Both the second and third trends are based in social psychology. The impact of operations research and the need to make important decisions quickly seems to have triggered the particular fruition of Delphi.

Of the first two trends, Weaver suggests that "An exhaustive check of the documentation shows two disparate and neatly isolated traditions of research--the experimenters in the one tradition not citing studies in the other" (1970:4).[2] In addition, he observes that there are differences in these two traditions: The individual trait researchers have tended to ignore the experimental task; the forecasting experimenters have focused on procedures, not traits.

Delphi: The Antecedents

Four studies by Nathan Israeli in the 1930's serve as the starting point for any discussion of modern forecasting methodologies. The first two concerned an experiment in which college students ranked the relative importance of the past, present, and future (1932a, 1932c). An overwhelming majority of students felt that the present and future were more important than the past and that "the present and future are almost of equal importance . . ." (1932a:489); 94.5% of the students said they were hopeful about the future.

Another experiment "dealt with wishes concerning future events assumed to be improbable" (1932a:489). Subjects responded to a series of questions about "the most important things which they wanted to see happen but which they were most certain would never happen" (1932a:489-490). Future events which seemed improbable ran the gamut from no more wars

[2]There is no implication in this discussion that Delphi responded continuously. Rather, its development seems to have been a fortuitous accident. This review is meant to trace and outline its precursors.

to smoking for females.[3]

Israeli then investigated the nature of the prediction of social, political, and scientific events by five groups of 607 college freshmen and sophomores (1933a). Participants were asked to pick from among 10 different events the most certain, most likely, or most probable. Examples of the type of question asked included the number of home runs Babe Ruth would hit in 1931,[4] how far ahead the weather can be predicted, the U. S. population at the end of the century, and the date of a cure for cancer. Participants behaved very much alike with uniformly appearing distributions; no between-group analysis was performed, however.

Israeli then examined "the notions students have of the possibilities and probability of a future decline of Western Civilization" (1932b:92), among 151 subjects drawn from college students majoring in Psychology, Mental Hygiene, U. S. History, and Sociology.

The task, similar to that in MacGregor (1938), Kaplan, Skogstad, and Girschick (1950), and Helmer (1966b), was to identify factors leading to a decline, to estimate the likely consequences and probabilities and date of decline. Cited factors were ranked overall and by group. "Subjects who hesitated most in dating a future decline of the West were least certain of it" (1933b:98).

Israeli also set out ten designs that constituted an overall plan to the problem of "measurement of attitudes to and interests in the future"

---

[3]This question of desirability was explored in greater detail by McGregor (1938) and Weaver (1969a, 1969b).

[4]Although the study wasn't published until 1932, data were collected in the fall and winter of 1930-1931.

(1933c:181; cf. Israeli:1932b). These plans, naive in some cases and generally incompletely delineated, are striking for their anticipation of much of the subsequent research in the area. The most provocative of these designs included:

A. Judgment of future events and situations based on past situation--in which each of two groups is supplied or not supplied with data on past events and asked to "project the past into the future." Symmetrical or nonsymmetrical series of years of varying ranges could be used. (McGregor pursued this line of interest in 1938.)

B. Judgment of future criticism of past and contemporary eminent men--in which a list of outstanding men (and women, although Israeli is unclear on this point) is evaluated by a group of judges "according to what they think future standards will be" (1933c:185), and Israeli felt the period of time under investigation could range from 1,000 to 10,000,000 years.

C. Judgment of Remote Future events--in which Israeli suggested that "people tend to become less dogmatic and more liberal or imaginative in regard to the near future" (1933c:183). According to Weaver (1970:7), subsequent research in this area has yielded contradictory results.

D. Possible future developments in different fields in which group and individual assessments might be compared. Israeli asked, "Will experts in related fields be able to visualize better the future developments in Field (x) than experts in foreign fields. In other words <u>is there a relation between degree of concentration in a field and the ability to imagine, map out, or picture future developmental possibilities in that particular field?</u>"

(1933c:189; emphasis added) (Both McGregor (1938) and Cantril (1938) were to examine the periphery of this question which is still central to the framework of Delphi.)

E. Probability of certain developments in the future--in which the possibilities developed in (D) would have probability estimates attached to them for particular time periods.

In (D) and (E) lie two of the elements of Delphi. Regrettably they were not to be linked for some years. The richness of Israeli's prescient questions about the future and man's perception of it are not dimmed by the incompleteness of those questions.

McGregor (1938) investigated

> . . . the predeterminers of prediction; with the relevant attitudes, wishes, and beliefs that are already existent when the individual is asked to predict, and with the objective conditions that have been present in the immediate past in the environment of the predictor. (1938:180; emphasis in original)

Three hundred sixty-seven students from Dartmouth, Bennington, and Columbia "about equally divided between the sexes" were asked to make predictions about nine events "involving a social situation of current interest" (1938:151). The subject population also included 33 professional teachers "who because of their expertness might be expected to predict with greater 'objectivity' than the average student" (1938:195). Data analysis revealed that the expectation went unfulfilled, however.

The items to be predicted included Roosevelt's re-election, a change

in business conditions, and a European War. In questions with Yes-No responses, participants were asked to assess the probabilities of occurrence. Other items sought to assess the subject's attitudes toward the events, his/her desires concerning the event, and the subject's information level.

McGregor concluded that "the influence of subjective factors upon prediction is limited by the degree of ambiguity of the stimulus situation, but also that this influence is dependent upon the importance for the predictor of the issues involved" (1938:192) [emphasis in original]. He suggested that beyong ambiguity, importance, and the nature of information available to the predictor (not the amount), other determinants might be optimism, pessimism, cautiousness, and presence or absence of skepticism. Incidentally, "the majority prediction was correct in every one of the nine instances" (1938: 198).

A study by Cantril, also in 1938, extended McGregor's work by asking several questions: (1) Can predictions about the time of an event be as accurate as predictions about outcomes? (2) Are there differences in local forecasts as opposed to distant ones? (3) What differences might there be in immediate (now) and temporally distant events? (4) What is the influence of information or knowledge? (5) of attitudes? (6) What is the influence of certainty?

Cantril used a 15-issue instrument on which subjects were asked to indicate outcomes, dates of events, and his/her degree of certainty; a brief attitude test about some of the short-range issues on the instrument was also included. The sample, which Cantril assumed was not random, used

35 individuals from each of 11 occupational groups, such as bankers, lawyers, newspaper editors, economists, historians, and sociologists, who were chosen because they were "outstanding in their occupational field" (1938:368), and 35 Communist Party members and 80 laymen. The events for which predictions were asked ran the gamut from the Spanish Civil War (very short-range) to the breaking up of the British Empire (long-range).

Cantril concluded that "people are more sure of their judgments regarding the direction of the outcome of an event than they are concerning the date of that outcome" (1938:378). Uniformity of judgments is proportionally and partially due to the degree to which the "factors leading up to an event can be resolved into a consistent and meaningful pattern" (1938:386). Further, where there is no external structure to aid prediction, a "relevant internal frame of reference" will structure the stimuli and determine the prediction" (1938:387). If there is neither external nor internal structure, "predictions are difficult and expressed with little certainty" (1938:388). Finally, when a prediction is based on an internal frame of reference, ego involvement tends to preclude objectivity. In a comparison of academicians and so-called men of affiars, the former were not as certain about outcomes of events as the latter.

Israeli first framed the significant questions about thought and future, and McGregor and Cantril explored this relationship in similar ways.

Israeli has been criticized for lack of data about age, sex, intelligence, and education, for use of inadequate statistics and faulty categorization

among other deficiences (Kastenbaum, cited in Weaver, 1969).

Both Cantril and McGregor provided additional information about participants, but the data were still generally incomplete. Other differences between these last two studies included the following: (a) McGregor compared professional teachers to students and Cantril compared professors to businessmen. These comparisons were perhaps the first attempts to look at "experts" vs. "non-experts;" (b) the relation of idiosyncratic traits to the accuracy of prediction was the focus of interest for McGregor, but Cantril was unable to examine that question as only two of his 15 specified events actually occurred; (c) McGregor examined the chances of occurrence only (as the events were presumed to be short-range), and Cantril attempted to get at the date of occurrence.

Slightly more than a decade later Kaplan and his colleagues investigated the quality of expert predictions, how predictions might be improved, and the extent to which predictive reliability could be appraised beforehand. He enlisted 26 participants, comprised of statisticians, mathematicians, economists, business administrators, an office manager, a secretary, and a writer; eight of the subjects held doctorates.

Questionnaires contained 10-16 items drawn equally from the social and natural sciences. A categorical format was utilized, and 123 questions were asked (29 had been dropped for various reasons), resulting in just over 3,000 responses. Each question referred to an event, the occurrence of which could be confirmed or not with a 20-week limit, and each question was divided into four exhaustive and exclusive outcomes for which probabilities

were to be estimated. Totals of the four estimates were to add up to 100. In addition, space was available for comments on each item.

Respondents received a new instrument each week for 13 weeks and had 3 hours in which to complete it. Half of the respondents were divided into three groups: (a) an independent group which responded as usual; (b) a cooperative group which discussed the items together but continued to respond individually; and (c) a joint group which discussed the items and came to a joint estimate. Subjects rotated through the three groups.

In terms of evaluation, the respondents were more successful in making short-term predictions than longer range ones (the last was 20 weeks, the limit of the study).

In terms of predictive improvement, Kaplan had given subjects two subscales of the American Council on Education Cooperative General Culture Test, Form X; no intelligence tests were given. "Predictive ability. . . shows a significant positive correlation with knowledge as measured by these tasks" [page 102], Kaplan, Girshick, and Skogstad commented, but the effect was not very great, and the authors added that there might be some factor of "judgment independent of detailed knowledge" (1950:103).

The analysis of the group results showed that the joint group (making one estimate) had a mean success of 67%; the cooperative group achieved 62%; and individuals were 52% successful. There were no significant differences in predicting social or natural science events, despite the social science bias of the subjects, although the confidence level was high in estimating social events. An analysis was also made of the statements or

comments made for each item; justification (a logical reason for prediction) was clearly related to successful prediction. In fact, guesses were successful 40% of the time.

There are some problems with the study: Subjects rotated among treatments with unknown effect; events had to be selected which were going to occur within five months; there was no way of controlling for the complexity of the predicted event; and minimal data on characteristics of subjects were provided.

The effect of increasing or decreasing the number of alternatives is unknown, and leaving the construction of choices to participants might have generated substantially different data. Nonetheless, this is an important study, for it links the earlier research of Israeli and McGregor to the subsequent activity of Normal Dalkey and Olaf Helmer.[5]

Recent Research

In an extended series of experiments, Dalkey and his colleagues compared face-to-face and anonymous interactions as related to accuracy and the impact of feedback on the nature of estimation (Dalkey, 1968a; 1968b; 1969; Dalkey, Brown, and Cochran, 1969; Brown, Cochran, and Dalkey, 1969; Dalkey, Brown, and Cochran, 1970; Dalkey and Rourke, 1971).

Almanac questions with known answers were used. It was assumed that the subjects, college and graduate students at UCLA, would have a

---

[5]The so called "original Delphi" was performed in 1953, but not declassified until the early 60's. This situation accounts in a small way for the gaps in the reporting of this approach.

"large amount of miscellaneous information relevant to the answers" (Dalkey, 1968a:6). In some of the studies, control and experimental groups were used; in others, the iterative structure (i.e., feedback) allowed each group to be its own control. In all, 11 experiments with 5,000 answers to 300 questions made up the study.

In comparing face-to-face vs. anonymous interaction, an ABBA pattern was used. The face-to-face condition used a discussion leader chosen at random (with no known effect) to list all available information in reaching a group consensus. The anonymous condition involved four rounds of estimate, with group results via medians and quartiles fed back between rounds. "The basic result," Dalkey wrote, "was that the median response of the questionnaire group was more accurate in 13 cases, and the consensus of the face-to-fact group was more accurate in seven cases" (1969:22). Typically, the "anonymous" subjects showed a spread of estimates that narrowed considerably with iteration and feedback. (Campbell's study--to be discussed subsequently--examined this issue somewhat differently.)

The individual first round responses were normalized, and a logarithmic distribution was generated. It appeared to Dalkey that "subjects were thinking as much . . . in terms of the size of the answers, as they were about the precise magnitude of the answer" (1969:25). Martino suggested that lognormality shows Delphi to be "a well-behaved process which can be described quantitatively" (1970:358) (and accounts for the use of the median as the major statistic in the research).

By using smaller groups (three instead of five members) and giving subjects first round results before the groups were assembled, Dalkey found that more responses were more accurate after discussion, and more responses were less accurate.

Than Dalkey examined what effect learning might have. One question was considered at a time with a cycle of estimate. feedback, re-estimate. There was no evidence for group learning over a 20-question sequence.

Other forms of feedback were investigated, that is, participants making estimates outside the quartiles were asked to justify their responses in a comment. No evidence accumulated to show that the reasons helped. In fact, answers deteriorated in accuracy following Round Two. The impact of formulating reasons was perused, but no improvement was noted, and when answers were sought for related questions, as well as the primary one, the outcome was indecisive (1969:59). One sexual sidelight was that women were both less accurate and more likely to change their estimates between rounds (see Bart, 1972).

Dalkey tried to use confidence estimates to identify a more knowledgeable subgroup which would be more accurate. "Although the higher the average confidence rating on a question, the smaller the group error" (1969:68-69), this notion was not confirmed.

In general, a wider spread of estimates tended to narrow (converge) with iteration and feedback, and the group response (median of final individual responses) became more accurate.

The group norm is more powerful than the hidden true answer. It

made no difference whether the group moved toward the true answer, for it consistently and independently moved closer to the median.

Dalkey did not address the question of why subjects changed their responses, and because of his interest in examining accuracy, did not look at any predisposing personality traits.

In a separate set of experiments, Dalkey, Brown, and Cochran used 20 almanac questions with 16 groups of 20 persons. By using self-rating information, the investigators were able to achieve "a significant improvement in the effectiveness of Delphi procedures . . . to select more accurate subgroups" (1969:18).

In 1966, Brown and Helmer examined the effect of feeding back medians and interquartile rages for almanac items to participants and examining the potential for a formula which would generate a more reliable estimator than the group median.

Twenty-three Rand researchers were asked to estimate answers and rate their expertise for 18 World Alnanac questions and two queries requiring computational effort. In Round Two, medians and interquartile ranges were fed back, as well as a frequency distribution of self-ratings. If the subject was outside the interquartile range, he/she was asked to justify the response; all subjects were asked to reconsider and re-rate Round One estimates. Round Three summarized the elicited reasons, and subjects were invited to offer criticisms.

The results showed that opinions converged, particularly through the pressure of having to justify a response. In Round One, 13 of the 20

interquartile ranges contained the correct answer, buy by Round Four, only seven did. On the other hand, the medians showed a tendency to move in the right direction, that is, toward the correct answer. By using the "most confident" third of the subjects, Round Four data showed a marked improvement for those subjects in comparison with the rest of the group. And in a finding parallel to those other experiments in the series, women rated themselves consistently lower in the experiments.

The last in this extended series of experiments compared the effect of different feedback on improvement in response to the customary sequence of 20 almanac questions with a number of variations.

The first sub-experiment involved feeding back medians and interquartile group responses to the control group, while experimental subjects received individual results. There was no difference in either the number of improved responses or the amount of improvement between the two groups.

Subsequently, Dalkey et al. ran this design again with one difference: the experimental group received one relevant fact for each question. "The result was decisive improvement in the experimental group in both the number of questions showing improved accuracy and in the amount of numerical improvement" (1970:v). This was the first experiment to reach statistical significance, and Dalkey et al., concluded that the nature of the feedback was not as critical as its quality--that is, it should offer an accurate resume of the group's behavior. Dalkey suggested elsewhere (1972:29) that convergence occured because of social pressure rethinking the problem and or transfer of information during feedback. No research followed, however. Once

again, females were less accurate initially, "but they are able to make as good, or better use of both feedback and feed-in to improve their estimates" (1970:vi).

It is important to remember that in all these studies, the answers, although obscure, were verifiable. There is no instance of the utilization of a future time period in any of the experiments. As the focus of Dalkey's interest, and that of his colleagues was on accuracy, it would obviously be imappropriate to utilize methodologies dependent on time consideration other than the past or present. It does not seem an untenable position to be concerned about some of the differences that might exist when different time periods become the object of interest.

Campbell compared "the forecasting accuracy of groups utilizing currently accepted direct confrontation methods to that of groups administered. . . the Delphi process" (1966:v).

Subjects were students in two graduate seminars in business forecasting, and each class was split into Delphi and interaction sub-groups.

Each group was "given the task of forecasting 16 economic series for the First Quarter 1966 . . ." (1966:32). Included were such measures as Gross National Product, Dow Jones 30 Industrials, Personal Savings, and U. S. Balance of International Payments.

Subjects were asked to rate their self-confidence in two ways--by assigning each of the 16 estimates to three "confidence groups" and by evaluating confidence in relation to his peer group.

The Delphi group was asked to make estimates for the 16 series

and for self-confidence in Round One; Round Two returned medians and interquartile ranges, asked for re-estimates, and justifying statements if the subject was outside the range. Round Three was the same as Round Two, except the reasons for "deviance" were summarized. Subjects were invited to criticize the deviant statements if they were so inclined. The last round allowed a final re-estimate of the series and self-confidence measures.

The interaction group forecast the same series under the constraints of limiting their activity to "fact finding, opinion appraisal, and the search for other relevant sources of information" (1966:46). Interaction subjects also rated their self-confidence as the Delphi condition participants had.

The results showed that the so-called Delphi group "forecasted more accurately, as a group and as individuals, than did group participants functioning under the controlled-interaction method" (1966:165). Further, Delphi participants improved their forecasts and showed decreased dispersion. Like Brown and Helmer's finding (1966), Campbell concluded that self-confidence ratings appeared to identify a more accuract group, but choosing the most confident did not mean choosing the most accurate.

Campbell also explored personality variables by administering personality scales and correlating these to performance measures during forecast discussions. The FIRO-B scale evaluated inclusion, control, and affective needs; the CAD evaluated aggression, compliance, and detached motifs. Campbell concluded--quite tentatively--that personality relationships to group performance existed, especially for those with affection and inclusion characteristics, as these could be persuaded to alter their views

in the discussion groups. At the same time, Delphi participants with the same level of affection and inclusion needs conformed to a significantly greater degree than other participants.

Two other related studies focused on personality variables and were performed by Weaver (1969) and Waldron (1970).

Weaver explored the variable of conceptual level as it related to the task of predicting.[6] He hypothesized that "complexity in interpersonal and social domains will be predictive of complexity in making judgments about the future" (1969:10; emphasis in original).

Employing the Paragraph Completion Test to measure integrative complexity, Weaver used 197 graduate students at Syracuse University who were randomly assigned to one of three conditions of "Delphi," each involving 16 events.

The first condition involved a range of 30 years for each event with an "expert date" provided based on a study by Helmer (1966b). Treatment Two was the same as the first, but the expert date was deleted. In both cases, subjects were informed that half the events were likely to occur by the end of the century, and the other half would probably occur in the next century.

In the list of the three conditions, subjects were not told that half the events were different from the other half, nor were they given the choice of a range of years as they had been, nor were they given expert dates.

---

[6]Low integrative complexity was equivalent to low conceptual level.

Additionally, all subjects were asked to rate his or her feelings of desirability if the event should occur and his/her familiarity with the event.

Weaver found that integrative complexity does predict performance. Integratively simple persons narrowed estimation ranges more than integratively complex persons did. Integratively simple persons widen ranges significantly when external cues are removed, and the convergence effect is independent of "willingness to assign variables which deviate from a norm" (1969:210).

Weaver suggested that the conditions are analogous to the Delphi rounds, an analogy which may be true if one considers Delphi rounds as independent.

The judgments of high integratively complex persons "correlated significantly with their self-rated feeling of 'desirability' regarding the events" (1970:19). This relationship did not hold for persons of low integrative complexity.

Weaver concluded:

> If complexity is characteristic of the expert (and it may well not be), his estimation of future dates may be related closely to his feeling of desirability with regard to the event in question. Such a relationship appears less likely to occur among the less stable but more concrete forecasters as suggested by this study. The complex person seems to be no more 'objective' (and perhaps less so) than concrete persons in forecasting future events. This is of course contrary to prior assumptions. (1970:19)

Waldron replicated Weaver's experiment with one major alteration: subjects were not confined to a one round paradigm which might be analogous to multiple Delphi rounds. They participated in three rounds of Delphi. Waldron also picked up a thread left loose by Dalkey in exploring time delay as a variable of environmental complexity.

One hundred eighteen graduate students in education were given the Paragraph Completion Test, and based on the results, groups were categorized as either HIC (high integrative complexity) or LIC (low integrative complexity). Members of both categories were randomly assigned to one of two treatments. The first of these involved one session with three Delphi rounds, in which subjects were asked to make earliest and latest estimates for an event (as in Weaver, supra). Feedback, received five seconds later, was controlled by randomly adjusting expert estimates between one and ten years. Estimates were revised, followed by controlled feedback, and a final re-estimation. The second treatment involved three consecutive afternoon meetings of 15 minutes during which one full round of estimation and feedback was completed.

Like Weaver, Waldron concluded that LIC's will be more variable and changeable than HIC's. Immediate feedback produced a different set of estimation range scores but did not appear to affect the tendency to change, especially for HIC's. Further, regardless of the effects of differential time delay, LIC's demonstrated a high tendency to alter their estimates than did HIC's.

Then the studies addressed events which had not yet occurred, unlike Dalkey's research discussed earlier. Campbell's time period was quite

short--just a few months and he was concerned with the level of an estimate, as time was a constant.

Weaver and Waldron, on the other hand, required estimation of earliest and latest dates, a decidedly different kind of task.

As Weaver observes, "The three studies did attempt to explain why some people conform to the group norm while other do not" (1970:20; emphasis added). Dalkey, et al. (1969, 1970, 1971) did not address, or attempt to address, idiosyncratic characteristics as they related to estimation as these studies do.

It seems apparent that a perusal of social-psychological attributes is important in trying to comprehend both the process and outcome of Delphi, although that is not always possible.

In reviewing these last two studies, the line between research into either accuracy or personality attitudes crosses that of an earlier direction, that is, the utilization of Delphi in forecasting technological changes by using "experts."

The Use of Experts

It is easy to trace the geneology of a methodology in a linear fashion, but that rarely reflects reality. Separating these lines of inquiry has been done only to permit them to be seen more clearly. Having pursued the aforementioned line of accuracy and personality traits, we must turn now to a consideration of experts, and in doing so examine the first "modern" Delphi study by Olaf Helmer.

In Helmer and Rescher's "On the Epistemology of the Inexact Science" lies the balance of the rationale for Delphi (1959). The advantages of feedback and controlled indirect interaction appear reasonable to this point. Now the expert is the last part of the relevant machinery. The expert, they argued, "has at his ready disposal a large store of (mostly inarticulated) background knowledge and a refined sensitivity to its relevance, through the intuitive application of which he [or presumably she] is often able to produce trustworthy personal probabilities regarding hypotheses in his area of expertness" (1959:38). If one wishes to explore issues related to the future, it appears intuitively reasonable to utilize experts. However, earlier studies (McGregor, 1938; Cantril, 1938) suggested that perhaps the matter was more in doubt than intuition would suppose. In any event, Helmer and Rescher believed that feedback would compensate for "considerations they might through inadvertence have neglected, and to give due weight to factors they were inclined to dismiss as inimportant on first thought" (1959:47).

Weaver felt that the use of experts caters "to the power structure, not the disenfranchised" (1972:21). Weaver's point is well taken: radical thinkers are often not perceived by the system, however construed, as experts on that system, and are excluded from participation. In the same way, experts are unlikely to be able to foresee fortuitous accidents, much less harmful ones.

On the other hand, if experts in an imaginary Delphi represent the point of view of those who hold the most powerful positions in whatever

field is being investigated, it is not unlikely that their opinions shall come to pass as they have the power to make it so. There is some disagreement over (a) the use of experts in Delphi and (b) the definition of an expert.

Insofar as this latter characteristic is concerned, Helmer and Rescher suggest that beyond knowledge, the measurement of which is left undetermined, an expert must have reliability and accuracy, aspects of behavior which imply prior experience with experts before a Delphi population is drawn. Nowhere in the literature does this seem to be the case. Helmer says quite simply, "Select your experts wisely" (1967b:4). He does not say how this might be done.

Finally, it is important to distinguish between judgments concerning fact (as in the Dalkey studies discussed above) and judgments of value on the one hand and future events on the other. Again, it is intuitively appealing to rely on expert opinion in these latter explorations. But recent studies (Weaver, 1969; Waldron, 1970) throw a shadow on such an assumption, for if personality traits are critical, to what extent can it be assumed that an expert is more objective? It is not necessarily so. Nonetheless, the role of the expert is typical and important in Delphi studies.

Technological Events

The last Delphi thread is concerned with technological events primarily and brings us to the original Delphi, carried out in the early 1950's but not unclassified for a decade (Helmer and Dalkey, 1962). Its purpose was to "apply expert opinion to the solution, from the viewpoint of a Soviet strategic planner, of an optimal U. S. industrial target system and

to the estimation of the number of A-bombs required to reduce the munitions output by a prescribed amount" (1962:1). Seven experts from four areas germane to the question participated in five weekly rounds in the following way:

Round One - asked primary question

Round Two - used results of interviews to generate subsidiary questions

Round Three - Reconsideration of original question and asked several subsidiary questions

Round Four - Additional information on target system, stockpike, etc., given

Round Five - Final bombing estimates made.

In Round Three where feedback was given, median estimates were used. In Round One, the estimates ranged from 50 to 5,000 A-bombs; by the last Round the range had decreased to 167 to 360 bombs. The median number of bombs decreased from 255 to 200 over the three estimation rounds.

Helmer and Dalkey reasoned that convergence "is almost inevitable in view of the progressively more penetrating analysis of the problem, achieved partly by means of the procedural feedback . . ." (1962:3). They did not explore either why the shifts occurred, the relation of personality characteristics to those shifts. Beyond that, the utilization of additional information clouded the issue somewhat, and freedom from experimenter bias was, in the view of the authors, not achieved, a situation which appears frequently in the literature.

A few years later, Gordon and Helmer (1966) used Delphi to survey longer range events in the areas of scientific break-throughs, population growth, automation, space progress, probability and prevention of war, and future weapons systems. Eighty-two (of 150) persons responded to one or more of four sequential instruments in the typical Delphi pattern. Medians and quartiles were the feedback statistics, and subjects taking "extreme" positions were asked to provide a reason for their position. Opinions converged in all areas, although the last group of estimates still demonstrated a substantial spread.[7] The authors felt that panel instability may have weakened convergence tendencies, that two months between rounds was too long, and that ambiguous questions may also have diluted some of the results. Concern was also expressed about undue averaging as biasing against the extreme predictor and about participants' general competence.

Gordon and Helmer concluded that "there is evidence . . . that the questions were thought provoking to many of our respondents, who may have found some reward for their labor through the mental stimulation to which the experiment exposed them" (1966:94), a line of belief expressed throughout much of the Delphi research.

Helmer reported briefly on another experiment (1967b) in which 20 UCLA Business School Faculty forecast ten economic and business indices over a four round Delphi which included a self-competence rating. Instead of using a group median, the responses of the subjects, the estimates of the

[7]The substantive outcome of the study is not discussed, as the object of interest here are the methodological aspects.

highest rated on competence were used as the "select" median. The select median was closer to the true value in just over two thirds of the cases.

In 1971, Dalkey and Rourke used Delphi, not to explore almanac type questions as had been done before, but to explore questions of value. Eighty upper class and graduate students at UCLA generated lists of value categories for higher education and the quality of life. Through a complex rating and grouping technique, 15 educational and 13 quality of life clusters emerged. In general, four groups using different rating methods in the typical Delphi sequence showed convergence patterns "comparable to similar indices for factual judgments" (1971:viii). The authors felt that "the outcome of these experiments appears to be that the Delphi procedures . . . are appropriate for generating and assessing value material" (1971:6).

Weaver's concise summary of the research in 1970 remains the clearest assessment of Delphi explorations today:

(a) Attitude, desirability, and ambiguity are related to assigning dates and probabilities to future events (exactly how remains unanswered).

(b) Distinguishing 'successful' predictors from unsuccessful predictors remains an unfulfilled, though desirable goal of some researchers.

(c) On the average, independently made forecasts on short-term events (less than one year) tend to be as accurate as group forecasts, and persons working in group discussions tend to produce more accurate estimates than when working alone.

(d) There is a slight tendency for controlled iterative processes to produce more 'accurate' estimates of short-term economic trends, and responses to alman-type questions than are produced in discussion groups.

(e) When people share their predictions with each other, the predictions tend to become more alike with the convergency seemingly taking place independent of the 'true' answer.

(f) People tend to assign higher probabilities to events when the conditions which surround the potential event are known and are unambiguous.

(g) People tend to be more certain about the likelihood that an event <u>will</u> occur sometime than just <u>when</u> it will occur.

(h) Relatively near events tend to be easier to predict with accuracy than relatively longer term events . . .

(i) People who can cite grounds or assumptions that support forecasts tend to be more accurate than others on short-term forecasting.

(j) Just precisely why convergency occurs is not yet clearly understood, but the direction of the evidence points to certain complex interrelationships between personality and the task conditions.

(k) Objectivity, i.e., disallowing one's feelings of desirability, in the predictive process seems to be a weak assumption regardless of whether the population is or is not 'expert.'

(1) A controlled iterative process using anonymous respondents, that is, the Delphi method, is no more likely to be free of muddling psychological variables than group discussions. (1970: 26, 27, 28 [abridged])

Delphi has been used in many areas, notably technological forecasting and education. One of the broadest studies to date was performed by deBrigard and Helmer in 1967 and concerned societal developments between 1970-2000. Another by Thompson (1973) concerned drug use and prevention, and Schneider looked at regional planning problems with Delphi (1972). Studies concerning facilities planning, the future of the family, among a host of others, have been published recently.

By definition Delphi is amenable to and appropriate for the collection and revision of opinion in virtually any field, but perhaps it is too early to accept the methodology without question.

Delphi and Education

In an article concerned with Delphi in higher education settings, Judd outlines three major areas in which the technique has been used (1972): educational goals and objectives, curriculum and campus planning, and effectiveness and cost-benefit measures. These areas are not limited to higher education, as many studies at different levels and in other educational areas have been performed. In fact, there are undoubtedly hundreds of such investigations in all sorts of forms and conditions but labeled as Delphi.[8]

[8]Delphi also had a substantial impact in corporate and public settings. See Business Week (1970), Fusfeld and Forster (1971) for examples. A major Delphi by the National Industrial Conference Board and the Opinion Research Corporation (undated) concerning problems for managers in the coming ten to twenty years is a good example.

In this section, eight studies will be reviewed within the three areas delineated by Judd. There is no contention that these are studies of ultimate rigor; in general, however, they represent typical research patterns.

Educational Goals and Objectives

In 1965, Helmer participated in an Educational Innovations Seminar at the Institute of Government and Public Affairs at UCLA.[9] Carrying on in the explorations of value rather than fact, Helmer thought Delphi to be applicable "to all phases of educational planning, at the federal, state, local, or individual institutional level" (1966:6). In this study, he wanted to use a panel of experts to establish appropriate educational goals, collect suggestions of potential innovations, modify the innovations in comparison with the goals, and estimate dollar cost. Over one hundred persons participated as seminar members, conference participants, or outside experts.[10]

Group A suggested innovations and, after they were collated and edited, evaluated them in terms of their importance, desirability, and feasibility. The innovations, totalling 93, were grouped under eleven headings. The group was to allocate an imaginary 5-year budget of 10 billion dollars among the proposed changes. Innovations ran the gamut from Educational Research and Development to New Equipment, and Adult Retraining, among others.

Group B assigned the developed items to one of four cost categories:

---

[9]The seminar is reported in full in Hirsch and Colleagues (1967).

[10]See Adelson for a discussion related to this research.

essentially free, low cost (1-10 million dollars), medium cost (10-100 million dollars), and high cost (100 million dollars and up). Group C had much the same task as Group B, except it was divided into two groups and then assigned into one of four additional groups. These groups had to allocate 9 billion (High), 900 million (Medium), 90 million (Low), and 10 million (Essentially Free) dollars across the eleven categories.

Results showed the High Group emphasized increases in student participation, raising teachers' salaries, buying new equipment, and retraining adults.

The "Medium" group valued facilities, school administration, and methods and organization of instruction; the "Low" group also emphasized the administration and instructional aspects that the "Medium" group had, plus educational research and development. The "Essentially Free" group emphasized staff training and curriculum, along with the instructional variables of the "Lows" and "Mediums."

Participants did not focus on a few innovations but selected a broad range. Next to raising salaries, the next largest number of dollars was allocated to increasing student participation.

In general, all respondents approved of all items described under "Reorganization of Instruction and Program." Several items were also rejected by all three groups, including subsidization of private schools, reversal of larger school trend, and subsidizing on-the-job training.

Helmer does not say much beyond the observation that the study was merely a pilot. The use, if any, of the information is unclear, but this may be a typical finding for educational Delphi studies.

Norton (1970) used Delphi to "acquire the broadest possible exposure to the views of persons who, because of their geographical proximity, vocational responsibility, personal interest, or professional understanding might provide insights of value to the planning staff for Governors State University" (1970:2, emphasis added).

At the outset, it is imperative to observe that Norton's study was not confined to experts, nor was it focued on the future--that is, the longer range future.

He used three questionnaires, the first of which was to seek a "diversity of opinion," and the last two to "seek the convergence . . . of opinions" (1970:2).

The usual Delphi design was modified in two ways: The assembling of respondent lists was quite complex, as it initially involved 1,185 individuals in over 30 subgroups. Secondly, data fro. a similar study at another state university suggested that two mailings might be sufficient.

Norton received 315 responses to the first mailing, a response rate of 26 percent. The 50 items suggested by participants were parallel to items in the Institutional Functioning Inventory which was used--as a kind of check on the items developed by participants.

The second instrument was mailed to a revised list of respondents with 45.4 percent response rate. Items were analyzed by ranks within categories.

As a study of educational goals, Norton seems to have been involved in a study of educational goal legitimation. "Experts," however construed, were not the focal population, and opinion convergence was not measured because of the "one-shot" design. Norton describes the study as "Delphi-like" (1970:1). Perhaps more aptly it is a questionnaire under the Delphi guise.

Clarke and Coutts focused on the year 2000 in a study of the future of teacher education (1971). Forty of 55 chief administrative officers in the English language teacher education institutions in Canada participated.[11] Four rounds involved formulating statements about teacher education based on 20 general statements of the future of education, estimating probable dates of occurrence of the collated and edited statements, using feedback of results for re-estimation, and, in the final part, estimating the date of occurrence of the event in respondent's own institution.

The authors found some sharp differences between rounds and assumed that feedback produces rethinking. As one might expect, participants tended to estimate an event's occurring in their own setting at a time later than in teacher education settings generally.

A continuing question for this study, and others in education is--or should be--a familiar one by now: Does the study use experts or merely knowledgeable persons? Whatever difference may exist in the quality and character of those groups is not clear, but it seems probable that there are differences. One may be driven to the justification that knowledgeable--that is, powerful or well-placed--individuals make policy regardless of the presence or absence of expertise. It is also not clear what effect the year 2000 may have had on results. Presumably, respondents are less confident for estimates so many years in advance.

Huckfeldt used Delphi to "study the changes that would be likely to

---

[11] A related Delphi concerning future responsibilities of English teachers was reported by Farrell. (1971)

occur in postsecondary education during the next five to fifteen years" (1972:11). This was a large-scale Delphi and involved 525 potential participants from areas directly and indirectly related to higher education of whom 385 eventually participated. The five Delphi rounds covered five months and saw the response rate drop from 80 percent in the first round to 47 percent in the last.

Huckfeldt's pattern by round was somewhat different from Delphi studies discussed thus far. After the generation of statements in Round One, subjects estimated the impact and likelihood of the change in Round Two. Additional statements were generated in the third Round, and Round Two results were fed back with the Round Two questions. Round Four repeated what was sought in Rounds Two and Three and asked in what time period the listed changes would occur. The final Round repeated the Round Four questions and added two new ones: Should the change occur, and who will most affect the change?

Huckfeldt concluded that respondents had difficulty in thinking about change beyond a five-year period, although agreement by the panel was substantial in most areas. In general, the study suggested that "changes in planning and management are most likely to occur and that the educational structure is least likely to change" (1972:22).

Huckfeldt's Delphi procedure was quite complex; perhaps the diversity of task for some respondents may have been appealling and stultifying for others, and the five rounds may have caused participant fatigue (as evidenced by the low return rate). That participants were mailed every round's

instrument whether they had completed the previous round may contradict any capacity to measure convergence as a function of individual changes in responses.

Another large scale Delphi reported by Reynolds (1973) relied on statements that were generated not by participants but by the author, various committees and other "experts." This change in procedure would appear to remove the study from Delphi and locate it with sequential questionnaires.

Cyphert and Gant sought to clarify the "opinions of clientele concerning goals for the School of Education" at the University of Virginia (1970:6). They arbitrarily defined the power structure as it "relates to teacher education in general and the School of Education in particular" (1970:6). Their population of 421 included college and University faculty, students, deans, statewide educators, persons with political influence, representation of the press, labor, Chamber of Commerce, and NAACP, and a sample of teacher educators chosen for their national repute.

Subjects were asked to suggest "prime targets" for the School of Education and their 751 statements generated by 298 (68 percent response rate) participants were reduced to 61 generic statements by five knowledgeable (not defined) persons acting as a Project Advisory Committee.

Round Two asked for priority estimates on a 5-point scale and included one bogus item to test for the capacity of the methodology to mold opinion. (It showed that, in fact, opinion could be molded.) Round Three reported both group modal results, as well as the individual's prior rating. If an individual differed from group priorities, he/she was asked to justify

his/her departure. The last round, to which 62 percent responded, asked for final ratings.

The authors concluded that the bulk of opinion change occurred after one round of feedback--that is by the third questionnaire. Administrators saw Delphi as an acceptable way of establishing institutional goals; teachers and other employees perceived the development of institutional goals as interferring with their own personal goals.

This Delphi, like the others reported, examined events thought to be desirable as distinguished from events thought to be likely. As Weaver suggests (1972b:31-32), one cannot discern what difference might exist in the response set of participants in addressing either type of issue.

Such studies may evoke not only notions of legitimization for events of a certain kind but also measures of potential resistance or hostility to those events. To this degree they are a participant-developed projective test for organization goal setting or change structuring. In a strict sense, they resemble Delphi.

Curriculum and Campus Planning

Uhl (1971) investigated the present and desirable institutional goals from the point of views of a number of on- and off-campus groups through the use of Delphi centered on a preliminary form of the Institutional Goals Inventory (IGI). A task force modified the IGI and developed a 105 item instrument.

Five institutions representing black, white, large and small, college and university, and public and private colleges and universities were chosen

as the population base. Groups were chosen by a stratified random sampling procedure and included male and female students and faculty from each department, academic administrators, trustees and alumni. Other groups represented were local political, occupational, religious, and minority groups, plus newspaper editors and mayors. At two schools, parents were also sampled. Because of the large sample size of 1,000, the usual first step of generating statements was eliminated and the aforementioned 161 substituted. Respondents were asked to evaluate is and should be goals; space was provided for both comments and additional goals. In Round Two, only group modal responses were returned, not the individual's prior responses in an effort to reduce potential defensiveness. Round Four asked for re-estimates and made available minority opinions.

Unlike some of the other Delphi studies, Uhl's showed a very high rate of response, dropping from 85 percent to 75 percent over three rounds. That students were recompensed undoubtedly helped maintain a high rate.

In terms of convergence (in which Uhl was quite interested) both individuals and groups showed significant changes, and the process did not produce less accurate data, based on an analysis of group error.

Uhl pointed out the two major distinctions between his study and a conventional Delphi, that is, many participants could in no way be construed as experts and the use of the IGI was a major shift, although quite similar to Norton's study (1970). The major concern is whether, in fact, the study was a Delphi or simply resembled one in format.

Another in the consensus-seeking studies[12] was performed by Scannell (1972) who examined educational, lay, and political group opinion relating to skills developed in professional teacher education courses in Kansas. He employed competency statements provided by teacher educators in the state. Two hundred seventy-six participants went through a Delphi-like estimation, feedback, re-estimation process. In general, teachers changed their estimates least; professors and administrators changed the most, and most of the changes occurred, as in Uhl's study, after the first information feedback.

Effectiveness and Cost-Benefit Measures

In 1971, Fox and Brookshire attempted "to formulate a definition of effective teaching and further to determine the relative importance of the various elements of such a definition" (1971:1-2) at the University of Northern Colorado. Ninety-one faculty generated 195 descriptors, and in the next round grouped the items and suggested a name for the grouped characteristics. The top 50 checked items were submitted to participants in Round Three with a request to develop "natural" categories. Seven emerged (to which two levels of descriptors were attached, e.g., Interpersonal Relationships: Approachable, Fair, Honest), and these results became the base for fictional profiles and other kinds of analysis.

This small group of "educational" Delphis is reasonably representative in quality and breadth of interest of a plethora of other studies. Quite obviously, they lack a resemblance to many of the characteristics typical of

[12]An option recommended by Helmer in a paper in 1967 (c).

Helmer's work and some of Dalkey's, as well as that of Campbell, Weaver, and Waldron.

In general, the bulk of the educational studies

1. do not use experts, even broadly defined;
2. do not consistently employ item generation;
3. do not become concerned explicitly with the short- or long-range future;
4. are frequently susceptible to experimenter bias, particularly in the statement categorization phase;
5. do not provide personality or demographic data on subjects;
6. do not explore characteristics of nonrespondents;
7. do not treat interval between rounds as a potentially significant variable; and
8. do not examine either participant commitment or fatique (see 6 above)[13]

Beyond these relatively specific concerns, a number of more general issues are related to the appropriateness or validity of Delphi. The research is hardly conclusive, although perhaps it could be described as tantalizing.

Two recent critics who have addressed this question with vigor are Sackman (1974) and Weaver (1971, 1972a, 1972b).

Sackman held Delphi up against the basic standards of empirical science as represented by APA standards and found it lacking in a number of areas: Interpretive Standards, Empirical Validity, Standards for the Use

[13]Weatherman and Swenson (1974) discussed these issues in a recent article.

of Experts, Theoretical Standards, Reliability, and Experimental Sampling Standards.

1. Statistical significance is rarely reported either for precision of estimates or for mean of median differences.

2. Delphi "does not measure events themselves, nor does it incorporate systematic hypotheses and empirical feedback from such events" (1974:15).

3. No standards for the use of experts have either evolved or been developed, and Sackman found a contradiction in the independence of the first round judgment vs. the dependence of subsequent rounds through feedback and the inclusion of comments.

4. The theoretical standards of Delphi are quite low because baseline data is rarely made available, population characteristics are rarely reported, and "the reasons, theories, and hypothetical constructs of expert panelists are covert, rather than overt" (Sackman, 1974:22).

5. Delphi instruments lack reliability, and studies are rarely replicated. The expedient selection of panelists reflects low experimental sampling standards, and overall, "It should be abundantly clear that conventional Delphi neglects virtually every major area of professional standards. . ." (Sackman, 1974:27). Occasionally vitriolic and almost righteously indignant, Sackman concluded that Delphi (a) is not scientifically tenable; (b) generates a conformist-consensus, ergo, a specious one; (c) clouds the no-significant difference conclusions of years of individual versus group performance experimentation; (d) uses amorphous questions; (e) generates stereotyped

responses; (f) produces ambiguous and meaningless responses; (g) assumes experts really are experts; (h) can, through results, reflect vested interests primarily; (i) inhibits the adversary process; (j) is an unvalidated technique; and (k) should be dropped until it is shown to be scientifically tenable (1974:33-65 [abridged]).

There is a good deal of merit in Sackman's observations, but they do not resolve the issue of addressing the future more scientifically. Sackman contrasted the somewhat primitive state of Delphi with rather sophisticated and highly developed standards pertaining predominantly to past conclusions and not future predictions. Over a period of years, Delphi ideally may acquire increased scientific respectability or, at least, an aura of acceptance, which does not exist now, as the theory and methodology unfold and are developed.

Weaver (1972a, b) approached Delphi differently from Sackman. He observed that "Delphi, like the future it was intended to foretell, has not turned out to be what we expected (1972:34). The major weakness for Weaver is that Delphi lacks any explanatory quality. It may be plausible, but only on the basis of expert agreement, which, to him, is insufficient. A situation in which "one is left to find his (or her) way backwards to the future" (1972a:6) obviously lacks explanation. In a more general way, Weaver would undoubtedly not be too discomfited by Sackman's observations.

> Weaver's rationale for continuing to employ Delphi is that it in combination with other tools, is a very potent device for teaching people to think about the future of education in much more complex

ways than they ordinarily would. When this use of Delphi is understood it is found to be a useful instrument for something more important than what it is designed for, namely a general teaching strategy. What this means is that initially the way to get educators to make better decisions--decisions which account for alternative consequences--is to enhance their capacity to think in complex ways about the future, and Delphi seems ideally suited to such a purpose. Indeed, educators may find in Delphi and other forecasting tools a better pedagogy. (1972b:54)

It may still be too early to relegate Delphi with a half-hearted whimper to triviality. Surely, educators have done the methodology no great service through inadvertence and carelessness, as well as in trying to wheedle political ends by means of a methodology that seems to be reasonably rigorous.

Even with the sharp divergence of opinion about Delphi as a legitimate research methodology, there are perhaps two reasons for considering Delphi as a plausible research methodology: First, it is a convenient and relatively non-technical methodology which is appropriate for use with a population not primarily research-oriented. Secondly, it appears a convenient vehicle for exploring probable and possible futures of in what Strauch called "squishy" areas (1974). Put differently, as an exploratory methodology, Delphi is sufficiently flexible to generate a broad range of responses and at the same time offers a reasonable structure within which responses might be organized and evaluated by participants. To that end, Delphi continues to be a useful,

if by now somewhat tarnished, vehicle to assist researchers and participants alike in exploring certain parameters of the future.

REFERENCES

Adelson, M. Planning Education for the Future: Comments on a Pilot Study. American Behavioral Scientist, 1967, 10 (7), 1-31.

Ayres, R. Technological Forecasting and Long Range Planning. New York: McGraw-Hill, 1969.

Bart, P. Why Women See the Future Differently From Men. In A. Toffler (ed.), Learning for Tomorrow. New York: Vintage, 1972.

Beckett, S. Waiting for Godot. New York: Grove Press, 1970.

Brown, B., Cochran, S., & Dalkey, N. The Delphi Method, II: Structure of Experiments (RM-5957-PR). Santa Monica, California: Rand, June, 1969.

Brown, B., & Helmer, O. Improvement in the Reliability of a Consensus Through the Use of Self-Ratings. In O. Helmer, Social Technology. New York: Basic Books, 1966.

Campbell, R. A Methodological Study of the Utilization of Experts in Business Forecasting (Doctoral dissertation, University of California at Los Angeles, 1966). Dissertation Abstracts, 1967, 27, 3170A-3171A. (University Microfilms No. K67-4491)

Cantril, H. The Prediction of Social Events. Journal of Abnormal and Social Psychology, 1938, 33, 364-389.

Clarke, S. & Coutts, H. The Future of Teacher Education. Unpublished manuscript, University of Alberta, Canada, 1971.

Cyphert, F. & Gant, W. The Delphi Technique: A Tool for Collecting Opinions in Teacher Education. Paper presented at the AERA Symposium: Exploring the Potential of the Delphi Technique by Analyzing its Application. Minneapolis, Minnesota, March 4, 1970.

Dalkey, N. Experiments in Group Prediction (P3820). Santa Monica, California: Rand, April 1968. (a)

Dalkey, N. Predicting the Future (P3948). Santa Monica, California: Rand, October, 1968. (b)

Dalkey, N. The Delphi Method: An Experimental Study of Group Opinion (RM-5888-PR). Santa Monica, California: Rand, June, 1969.

Dalkey, N. Delphi. In J. Martino (Ed.), An Introduction to Technological Forecasting. London: Gordon and Breach, 1972. [Paper originally delivered in 1967.]

Dalkey, N., Brown, B., & Cochran, S. The Delphi Method, III: Use of Self-Ratings to Improve Group Estimates (RM-6115-PR). Santa Monica, California: Rand, November, 1969.

Dalkey, N., Brown, B., & Cochran, S. The Delphi Method, IV: Effect of Percentile Feedback and Feed-In of Relevant Facts (RM-6118-PR). Santa Monica, California: Rand, March, 1970.

Dalkey, N., & Rourke, D. Experimental Assessment of Delphi Procedures With Group Value Judgments (R-612-ARPA). Santa Monica, California: Rand, February, 1971.

deBrigard, R., & Helmer, O. Some Potential Societal Developments - 1970-2000. Middleton, Connecticut: Institute for the Future, Report R-7, April, 1970.

de Jouvenel, B. The Art of Conjecture, (N. Lang, trans.). London: Weidenfeld and Nicolson, 1967.

Farrell, E. Deciding the Future: A forecast of Responsibilities of Secondary Teachers of English, 1970-2000 AD. Urbana, Illinois: National Council of Teachers of English, 1971.

Forecasters Turn to Group Guesswork. Business Week, March 14, 1070, 130-134.

Fox, A., & Brookshire, W. Defining Effective College Teaching Using the Delphi Technique and Multiple Linear Regressions. Paper presented to AERA, New York, January, 1971.

Fusfeld, A., & Foster, R. The Delphi Technique: Survey and Comment. Business Horizons, June, 1971, 63-74.

Gordon, T., & Helmer, O. Report on A Long-Range Forecasting Study. In O. Helmer, Social Technology. New York: Basic Books, 1966.

Helmer, O. Social Technology. New York: Basic Books, 1966. (a)

Helmer, O. The Use of the Delphi Technique in Problems of Educational Innovation (P3499). Santa Monica, California: Rand, December, 1966. (b) [same as Helmer, O., 1967a]

Helmer, O. The Delphi Technique and Educational Innovation. In W. Hirsch (Ed.), and Colleagues, Inventing Education for the Future. San Francisco, California: Chandler, 1967. (a) [same as Helmer, O., 1966b]

Helmer, O. Analysis of the Future: The Delphi Method (P3558). Santa Monica, California: Rand, March, 1967. (b)

Helmer, O. Systematic Use of Expert Opinions (P3721). Santa Monica, California: Rand, November 1967. (c)

Helmer, O. Long-Range Forecasting--Roles and Methods. Paper delivered at the conference on Forecasting the Future at Harrowgate, England. Middletown, Connecticut: Institute for the Future, Report P-7, May, 1970.

Helmer, O., & Dalkey, N. An Experimental Application of the Delphi Method to the Use of Experts (RM-727-PR-Abridged). Santa Monica, California: Rand, July, 1962.

Helmer, O., & Rescher, N. On the Epistemology of the Inexact Sciences. Management Science, 1959, 6, 25-52.

Hirsch, W., & Colleagues. Inventing Education for the Future. San Francisco, California: Chandler, 1967.

Huckfeldt, V. Imaging Future Organization in Higher Education. In S. Popper (Ed.), Imaging Alternative Future Organizations. Minneapolis, Minnesota: University of Minnesota, 1972.

Israeli, N. The Psychopathology of Time. Psychology Review, 1932, 39, 486-491. (a)

Israeli, N. Wishes Concerning Improbable Future Events: Reactions To The Future. Journal of Applied Psychology, 1932, 16, 584-588. (b)

Israeli, N. The Social Psychology of Time. Journal of Abnormal and Social Psychology, 1932, 27, 209-213. (c)

Israeli, N. Attitudes To the Decline of the West. Journal of Social Psychology, 1933, 4, 201-222. (b)

Israeli, N. Measurement of Attitudes and Reactions to the Future. Journal of Abnormal and Social Psychology, 1933, 28, 181-193. (c)

Jantsch, E. Technological Forecasting in Perspective. Paris: Organization for Economic Co-operation and Development, 1967.

Johnston, D. Social Indicators and Social Forecasting. In S. Popper (Ed.), Imaging Alternative Future School Organizations. Minneapolis, Minnesota: University of Minnesota, 1972.

Judd, R. Use of Delphi Methods in Higher Education. Technological Fore-Casting and Social Change, 1972, 4, 173-186.

Kaplan, A., Skogstad, A., & Girshick, M. The Prediction of Social and Technological Events. Public Opinion Quarterly, 1950, 14 (Spring), 93-110.

Martino, J. The Lognormality of Delphi Estimates. Technological Forecasting, 1970, 1, 355-358.

Martino, J. K. Technological Forecasting for Decisionmaking. New York: American Elsevier Publishing, 1972. (a)

Martino, J. Methods of Technological Forecasting. In W. Hahn & K. Gordon (Eds.), Assessing the Future and Policy Planning. London: Gordon and Breach, 1972. (b)

McGregor, D. The Major Determinants of the Prediction of Social Events. Journal of Abnormal and Social Psychology, 1938, 33, 179-204.

National Industrial Conference Board. Perspectives for the 70's and 80's. New York: National Industrial Conference Board, undated.

Nisbet, R. The Year 2000 and All That. Commentary, 1968, 45 (6), 60-66.

Norton, D. Governors State University Needs Assessment Survey. Evanston, Illinois: Educational Testing Service, April, 1970.

Polak, F. Prognostics. Amsterdam: Elsevier Publishing, 1971.

Reynolds, M. Delphi Survey: A Report of Rounds I and II. Arlington, Virginia: Council for Exceptional Children, September, 1973.

Sackman, H. Delphi Assessment: Expert Opinion, Forecasting, and Group Process (R-1283-PR). Santa Monica, California: Rand, April, 1974

Scannell, D. The Use of the Delphi Technique To Gain Consensus on the Professional Education Components of Teacher Certification Requirements in Kansas. Final Report (Project No. 1GO72), National Center for Educational Research and Development, U. S. Office of Health, Education, and Welfare, Office of Education, Washington, D. C.

Schneider, J. The Policy Delphi: A Regional Planning Application. Technological Forecasting and Social Change, 1972, 3. 481-497.

Strauch, R. A Critical Assessment of Quantitative Methodology as a Policy Analysis Tool (P5282). Santa Monica, California: Rand, August, 1974.

Thompson, L. A Pilot Application of Delphi Techniques to the Drug Field: Some Experimental Findings (R-1124). New York: Rand, June 1973.

Uhl, N. Encouraging Convergence of Opinion Through the Use of the Delphi Technique in the Process of Identifying an Institution's Goals. Durham, North Carolina: Educational Testing Service, February, 1971.

Waldron, J. An Investigation into the Relationship Among Conceptual Level, Time Delay of Information Feedback and Performance in the Delphi Process (Doctoral dissertation, Syracuse University, 1970). Dissertation Abstracts International, 1971, 31, 5862A-5863A. (University Microfilms No. 71-11,000)

Weatherman, R., & Swenson, K. Delphi Technique. In S. Hencley & J. Yates (Eds.), Futurism in Education. Berkeley, California: McCutchan, 1974.

Weaver, W. An Exploration into the Relationship Between Conceptual Level and Forecasting Future Events (Doctoral dissertation, Syracuse University, 1969). Dissertation Abstracts International, 1970, 30, 5354A-5355A. (University Microfilms No. 70-10, 405) (a)

Weaver, W. The Delphi Method: Background and Critique: Report I. Syracuse: Educational Policy Research Center, 1968. (b)

Weaver, W. The Delphi Forecasting Method. Phi Delta Kappan, 1971, 2, 267-271.

Weaver, W. Delphi, A Critical Review. Syracuse University Research Corporation, Educational Policy Research Center, February, 1972. (a)

Weaver, W. The Delphi Forecasting Method: Some Theoretical Considerations. In M. Marien & W. Ziegler (Eds.), The Potential of Educational Futures. Worthington, Ohio: Jones, 1972. (b)